Making Tensor Factorizations Robust to non-Gaussian Noise

Eric C. Chi[*]
Department of Statistics
Rice University
Houston, TX 15213
echi@rice.edu

Tamara G. Kolda[†]
Sandia National Laboratories
Livermore, CA 94551-9159
tgkolda@sandia.gov

Abstract

Tensors are multi-way arrays, and the CANDECOMP/PARAFAC (CP) tensor factorization has found application in many different domains. The CP model is typically fit using a least squares objective function, which is a maximum likelihood estimate under the assumption of i.i.d. Gaussian noise. We demonstrate that this loss function can actually be highly sensitive to non-Gaussian noise. Therefore, we propose a loss function based on the 1-norm because it can accommodate both Gaussian and grossly non-Gaussian perturbations. We also present an alternating majorization-minimization algorithm for fitting a CP model using our proposed loss function.

1 Introduction

The CANDECOMP/PARAFAC (CP) tensor factorization can be considered a higher-order generalization of the matrix singular value decomposition [4, 7] and has many applications. The canonical fit function for the CP tensor factorization is based on the Frobenius norm, meaning that it is a maximum likelihood estimate (MLE) under the assumption of additive i.i.d. Gaussian perturbations. It turns out, however, that this loss function can be very sensitive to violations in the Gaussian assumption. However, many other types of noise are relevant for CP models. For example, in fMRI neuroimaging studies, movement by the subject can lead to sparse high-intensity changes that are easily confused with brain activity [6]. Likewise, in foreground/background separation problems in video surveillance, a subject walking across the field of view represents another instance of a sparse high intensity change [13]. In both examples, there is a relatively large perturbation in magnitude that affects only a relatively small fraction of data points; we call this artifact noise. These scenarios are particularly challenging because the perturbed values are on the same scale as normal values (i.e., true brain activity signals and background pixel intensities). Consequently, there is a need to explore factorization methods that are robust against violations in the Gaussian assumption. In this paper, we consider a loss based on the 1-norm which is known to be robust or insensitive to gross non-Gaussian perturbations [8].

Vorobyov et al. previously described two ways of solving the least 1-norm CP factorization problem based on a linear programming and weighted median filtering [14]. Our method differs in that we use a majorization-minimization (MM) strategy [9]. Like [14] our method performs block minimization. An advantage of our approach is that each block minimization can be split up into many small and independent optimization problems which may scale more favorably with a tensor's size.

Throughout, we use the following definitions and conventions. All vectors are column vectors. The transpose of the i^{th} row of a matrix $\mathbf{A}$ is denoted by $\mathbf{a}_i$. The *order* of a tensor is the number of dimensions, also known as ways or modes. *Fibers* are the higher-order analogue of matrix rows and columns. A fiber is defined by fixing every index but one. A matrix column is a mode-1 fiber and a matrix row is a mode-2 fiber. The mode-n matricization of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is denoted by $\mathbf{X}_{(n)}$ and arranges the mode-n fibers to be the columns of the resulting matrix.

[*]This work was funded by DOE grant DE-FG02-97ER25308.

[†]This work was funded by the applied mathematics program at the U.S. Department of Energy. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

The rest of this paper is organized as follows. The robust iterative algorithm is derived in Section 2. In Section 3 we compare CPAL1 and the standard CP factorizations by alternating least squares (CPALS) in the presence of non-Gaussian perturbations on simulated data. Concluding remarks are given in Section 4.

2 Majorization-minimization for tensor factorization

MM algorithms have been applied to factorization problems previously [12, 3, 5]. The idea is to convert a hard optimization problem (e.g., non-convex, non-differentiable) into a series of simpler ones (e.g., smooth convex), which are easier to minimize than the original. To do so, we use majorization functions, i.e., h majorizes g at $\mathbf{x} \in \mathbb{R}^n$ if $h(\mathbf{u}) \geq g(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^n$ and $h(\mathbf{x}) = g(\mathbf{x})$.

Given a procedure for constructing a majorization, we can define the MM algorithm to find a minimizer of a function g as follows. Let $\mathbf{x}^{(k)}$ denotes the k^{th} iterate. (1) Find a majorization $h(\cdot|\mathbf{x}^{(k)})$ of g at $\mathbf{x}^{(k)}$. (2) Set $\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}} h(\mathbf{x}|\mathbf{x}^{(k)})$. (3) Repeat until convergence. This algorithm always takes non-increasing steps with respect to g. Moreover, sufficient conditions for the MM algorithm to converge to a stationary point are well known [11]. Specifically, the MM iterates will converge to a stationary point of g if g is continuously differentiable, coercive in the sense that all its level sets must be compact, and all its stationary points are isolated; and the function $h(\mathbf{x}|\mathbf{y})$ is jointly twice continuously differentiable in $(\mathbf{x}, \mathbf{y})$ and is strictly convex in $\mathbf{x}$ with $\mathbf{y}$ fixed.

2.1 Solving the ℓ_1 regression problem by an MM algorithm

We now derive an appropriate majorization function for approximate ℓ_1 regression; this is subsequently used for our robust tensor factorization. Given a vector $\mathbf{y} \in \mathbb{R}^I$ and a matrix $\mathbf{M} \in \mathbb{R}^{I \times J}$, we search for a vector $\mathbf{u} \in \mathbb{R}^J$ that minimizes the loss $L(\mathbf{u}) = \sum_i |r_i(\mathbf{u})|$ where $r_i(\mathbf{u}) = y_i - \mathbf{m}_i^{\mathsf{T}}\mathbf{u}$. Note that $L(\mathbf{u})$ is not smooth and may not be strictly convex if $\mathbf{M}$ is not full rank. Therefore, we instead consider the following smoothed and regularized version to $L(\mathbf{u})$:

$$L_{\epsilon,\mu}(\mathbf{u}) = \sum_{i=1}^{I} \sqrt{r_i(\mathbf{u})^2 + \epsilon} + \frac{\mu}{2}\|\mathbf{u}\|^2, \tag{1}$$

where ϵ and μ are small positive numbers. In this case, $L_{\epsilon,\mu}(\mathbf{u})$ at $\tilde{\mathbf{u}} \in \mathbb{R}^J$ is majorized by

$$h_{\epsilon,\mu}(\mathbf{u}|\tilde{\mathbf{u}}) = \sum_{i=1}^{I} \left\{ \sqrt{r_i(\tilde{\mathbf{u}})^2 + \epsilon} + \frac{r_i(\mathbf{u})^2 - r_i(\tilde{\mathbf{u}})^2}{2\sqrt{r_i(\tilde{\mathbf{u}})^2 + \epsilon}} \right\} + \frac{\mu}{2}\|\mathbf{u}\|^2. \tag{2}$$

Both the loss $L_{\epsilon,\mu}$ and its majorization $h_{\epsilon,\mu}$ meet the sufficient conditions that guarantee convergence of the MM algorithm to a stationary point of $L_{\epsilon,\mu}$. Since $L_{\epsilon,\mu}$ is strictly convex and coercive, it has exactly one stationary point. Thus, the MM algorithm converges to the global minimum of $L_{\epsilon,\mu}$. After some simplification, the iterate mapping can be expressed as

$$\mathbf{u}^{(m+1)} = \arg\min_{\mathbf{u}} \left\{ \sum_{i=1}^{I} \frac{r_i(\mathbf{u})^2}{\sqrt{r_i(\mathbf{u}^{(m)})^2 + \epsilon}} + \frac{\mu}{2}\|\mathbf{u}\|^2 \right\}. \tag{3}$$

Let $\mathbf{W}^{(m)} \in \mathbb{R}^{I \times I}$ be the diagonal matrix with $(\mathbf{W}^{(m)})_{ii} = (r_i(\mathbf{u}^{(m)})^2 + \epsilon)^{-1/2}$. Then the minimization problem (3) is a regularized weighted least squares problem with a unique solution, i.e.,

$$\mathbf{u}^{(m+1)} = \arg\min_{\mathbf{u}} \left\{ (\mathbf{y} - \mathbf{M}\mathbf{u})^{\mathsf{T}}\mathbf{W}^{(m)}(\mathbf{y} - \mathbf{M}\mathbf{u}) + \frac{\mu}{2}\|\mathbf{u}\|^2 \right\} = (\mathbf{M}^{\mathsf{T}}\mathbf{W}^{(m)}\mathbf{M} + \mu\mathbf{I})^{-1}\mathbf{M}^{\mathsf{T}}\mathbf{W}^{(m)}\mathbf{y}. \tag{4}$$

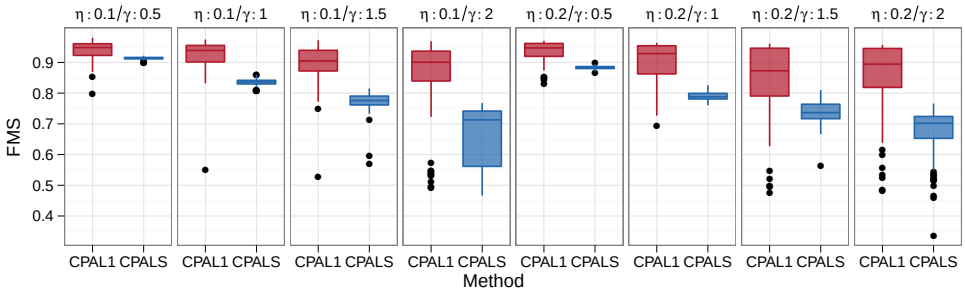
2.2 Tensor factorization using ℓ_1 regression

We now derive CPAL1 for a 3-way tensor $\mathcal{X}$ of size $I_1 \times I_2 \times I_3$ (it is straightforward to generalize the algorithm to tensors of arbitrary size). To perform a rank-R factorization we minimize $L_{\epsilon,\mu}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ which is the regularized approximate 1-norm of the difference between $\mathcal{X}$ and its rank-R approximation, where $\mathbf{A} \in \mathbb{R}^{I_1 \times R}, \mathbf{B} \in \mathbb{R}^{I_2 \times R}, \mathbf{C} \in \mathbb{R}^{I_3 \times R}$:

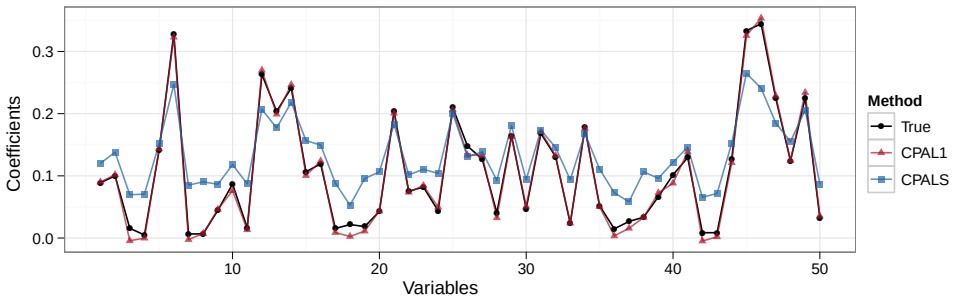
$$L_{\epsilon,\mu}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \sum_{i_1, i_2, i_3} \sqrt{\left(x_{i_1 i_2 i_3} - \sum_{r=1}^{R} a_{i_1 r} b_{i_2 r} c_{i_3 r} \right)^2 + \epsilon} + \frac{\mu}{2}\left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2 \right).$$

In a round-robin fashion, we repeatedly update one factor matrix while holding the other two fixed. Note that the mode-1 matricization of the rank-R approximation is $\mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}}$ where $\odot$ denotes the Khatri-Rao product [10]. Then the subproblem of updating $\mathbf{A}$ for a fixed $\mathbf{B}$ and $\mathbf{C}$ is

$$\min_{\mathbf{A}} \sum_{i=1}^{I_1} \sum_{j=1}^{I_2 I_3} \sqrt{\left(\left(\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}} \right)_{ij} \right)^2 + \epsilon} + \frac{\mu}{2}\|\mathbf{A}\|_F^2. \tag{5}$$



(a) The FMS distribution under different combinations of η and γ.



(b) A comparison of a single recovered factor column for a replicate when $\eta = 0.2$ and $\gamma = 2$. Here the FMS was 0.91 and 0.64 for CPAL1 and CPALS respectively. Factor columns were normalized for comparison.

Figure 1: Panel 1a shows on average that CPALS factorizations are sensitive to artifact noise. Panel 1b provides a close up comparison of the differences between the two methods for single recovered column in an instance where CPALS is less accurate.

This minimization problem is separable in the rows of $\mathbf{A}$, and the optimization problem for a given row is an ℓ_1 regression problem. Thus, we can apply the update rule (4) with $\mathbf{y}$ equal to the i^{th} row of $\mathbf{X}_{(1)}$ and $\mathbf{M} = \mathbf{C} \odot \mathbf{B}$. The other two subproblems are solved analogously.

3 Simulation Experiment

We compare the results of CPAL1 with CPALS implemented in the Tensor Toolbox [2] in the presence of Gaussian and artifact noise. We created 3-way tensors, $\mathcal{X}' \in \mathbb{R}^{50\times50\times50}$ of rank-5 as follows. We first generated random factor matrices $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{50\times5}$ where the matrix elements were the absolute values of i.i.d. draws from a standard Gaussian. The ijk^{th} entry of the noise free tensor $\mathcal{X}$ was then set to be $\sum_{r=1}^{R} a_{i_1 r} b_{i_2 r} c_{i_3 r}$. Then to each $\mathcal{X}$ we added dense Gaussian noise and artifact outliers. All random variables we describe were independently drawn. We generated an artifact tensor $\mathcal{P}$ as follows. A fraction η of the tensor entries was selected randomly. We then assigned to each of the selected entries a value drawn from a Gamma distribution with shape parameter 50 and scale parameter $1/50$. All other entries were set to 0. For the dense Gaussian noise tensor $\mathcal{Q}$, the entries q_{ijk} were i.i.d. draws from a standard Gaussian. The tensor $\mathcal{X}'$ was obtained by adding the noise and artifact tensors to $\mathcal{X}$:

$$\mathcal{X}' = \mathcal{X} + \gamma \frac{\|\mathcal{X}\|_F}{\|\mathcal{P}\|_F} \mathcal{P} + 0.1 \frac{\|\mathcal{X}\|_F}{\|\mathcal{Q}\|_F} \mathcal{Q}$$

for $\eta = 0.1, 0.2$ and $\gamma = 0.5, 1.0, 1.5$ and 2.0. For all combinations of η and γ the scaled values of q_{ijk} were less than the largest value of $\mathcal{X}$.

For every pair (η, γ) we performed 100 rank-5 factorizations under the two methods. For CPAL1 computations we set $\epsilon = 10^{-10}$ and $\mu = 10^{-8}$. Initial points for all tests were generated using the n-mode singular vectors of the tensor (i.e., the `nvecs` command in the Tensor Toolbox). To assess the goodness of a computed factorization we calculated the factor match score (FMS) between the estimated and true factors [1]. The FMS ranges between 0 and 1; an FMS of 1 corresponds to a perfect recovery of the original factors.

Figure 1a shows boxplots of the FMS under both methods. The scores for CPALS decreased as the contribution of non-Gaussian noise increased. In contrast regardless of the noise distributions applied CPAL1 tended to recover the true factorization with the exception of occasionally finding local minima,

Figure 1b compares one column of one recovered factor matrix when $\eta = 0.2$ and $\gamma = 2$ for the two methods. In this instance the CPALS factorization has some trouble recovering the true factor column. In this example the FMS was 0.91 and 0.64 for CPAL1 and CPALS respectively. The median CPALS FMS was about 0.7, so the example shown is somewhat typical. The factorization is not terrible qualitatively, but the errors in the Factor 2 estimates do fail to capture details that CPAL1 solution does.

4 Conclusion

We derived a robust tensor factorization algorithm based on an approximate 1-norm loss. In comparisons with methods using an LP solver we found that our method performed slightly faster on tensors of similar size to those factored in the simulation experiments of this paper (not shown). We suspect the performance gap may widen depending on the size of the tensor. Indeed, to factor an arbitrary tensor of size $I_1 \times \cdots \times I_N$ the LP update for the i^{th} factor matrix would be an optimization problem over $2\prod_{n\neq i} I_n + R$ parameters. In contrast, the i^{th} factor matrix update consists of I_i independent ℓ_1 minimizations over R parameters. Moreover, the independence of these minimizations present speed-up opportunities through parallelization.

Our simulations demonstrated that there are non-Gaussian noise scenarios in which the quality of CPALS solutions suffer while those of CPAL1 tend to be insensitive to the presence of non-Gaussian noise. In simulation studies not shown we have seen that not all non-Gaussian perturbations cause noticeable degradation in the CPALS factorization. Conversely, there are situations when CPAL1 struggles as much as CPALS in the presence of artifact noise, e.g. when the data tensor is sparse as well. We conjecture that CPAL1 is most suited to handle artifact noise when the data tensor is dense. Finding an alternative to the 1-norm loss for sparse data with non-Gaussian noise is a direction for future research.

References

[1] E. ACAR, D. M. DUNLAVY, T. G. KOLDA, AND M. MØRUP, *Scalable tensor factorizations for incomplete data*, Chemometrics and Intelligent Laboratory Systems. in press. Available at `http://csmr.ca.sandia.gov/~tgkolda/pubs/bibtgkfiles/CILS-CPWOPT-preprint.pdf`.

[2] B. W. BADER AND T. G. KOLDA, *Matlab tensor toolbox version 2.4*. `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/`, March 2010.

[3] R. BRO, N. D. SIDIROPOULOS, AND A. K. SMILDE, *Maximum likelihood fitting using ordinary least squares algorithms*, Journal of Chemometrics, 16 (2002), pp. 387–400.

[4] J. D. CARROLL AND J. J. CHANG, *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition*, Psychometrika, 35 (1970), pp. 283–319.

[5] I. DHILLON AND S. SRA, *Generalized nonnegative matrix approximations with bregman divergences*, Advances in neural information processing systems, 18 (2006), p. 283.

[6] K. J. FRISTON, S. WILLIAMS, R. HOWARD, R. S. FRACKOWIAK, AND R. TURNER, *Movement-related effects in fMRI time-series*, Magnetic Resonance in Medicine, 35 (1996), pp. 346–355.

[7] R. A. HARSHMAN, *Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA working papers in phonetics, 16 (1970), pp. 1–84. Available at `http://www.psychology.uwo.ca/faculty/harshman/wpppfac0.pdf`.

[8] P. J. HUBER AND E. M. RONCHETTI, *Robust statistics*, John Wiley & Sons Inc, 2009.

[9] D. R. HUNTER AND K. LANGE, *A tutorial on mm algorithms*, The American Statistician, 58 (2004), pp. 30–37.

[10] T. G. KOLDA AND B. W. BADER, *Tensor decompositions and applications*, SIAM Review, 51 (2009), pp. 455–500.

[11] K. LANGE, *Numerical Analysis for Statisticians*, Springer, 2010.

[12] D. D. LEE AND H. S. SEUNG, *Algorithms for non-negative matrix factorization*, Advances in neural information processing systems, 13 (2001).

[13] L. LI, W. HUANG, I. Y. GU, AND Q. TIAN, *Statistical modeling of complex backgrounds for foreground object detection*, Image Processing, IEEE Transactions on, 13 (2004), pp. 1459 –1472.

[14] S. A. VOROBYOV, Y. RONG, N. D. SIDIROPOULOS, AND A. B. GERSHMAN, *Robust iterative fitting of multilinear models*, IEEE Transactions on Signal processing, 53 (2005), pp. 2678–2689.